DOCUMENT RESUME

ED 457 194 TM 033 292

AUTHOR Capar, Nilufer K.; Thompson, Tony; Davey, Tim
TITLE Using Out-of-Scale Information To Increase the Precision of Test Scores.
PUB DATE 2000-04-00
NOTE 26p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 25-27, 2000).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Ability; *Adaptive Testing; *Computer Assisted Testing; Item Response Theory; Mathematics; Reading Comprehension; Scaling; *Scores; Simulation; Test Items

ABSTRACT

Information provided for computerized adaptive test (CAT) simulees was compared under two conditions on two moderately correlated trait composites, mathematics and reading comprehension. The first condition used information provided by in-scale items alone, while the second condition used information provided by in- and out-of-scale items together in computing the total information provided for simulees at their true ability levels. Information gains and associated bias and standard error measures were reported. Overall, information provided for simulees was higher for the second condition. Information provided increased approximately 17% for simulees on the reading comprehension and mathematics CATs when items were allowed to contribute information to both CATs (with their out-of-scale parameters). Results show that out-of-scale information can be used to improve a given measurement procedure without increasing the test length and testing time. For this reason, it merits further exploration. (Contains 6 figures, 2 tables, and 26 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 457 194

TM

# Using Out-of-Scale Information to Increase the Precision of Test Scores

Nilufer K. Capar

Florida State University

Tony Thompson & Tim Davey

ACT, Inc

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
N. Capar
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)



## Abstract

Information provided for CAT simulees was compared under two conditions on two moderately correlated trait composites, math and reading comprehension. The first condition used information provided by in-scale items alone, while the second condition used information provided by in- and out-of-scale items together in computing the total information provided for simulees at their true ability levels. Information gains and associated bias and standard error measures were reported.

Overall, information provided for simulees was higher for the second condition. Information provided increased approximately 17% for simulees on the reading comprehension and math CATs when items were allowed to contribute information to both CATS (with their out-of-scale parameters). Results show that out-of-scale information can be used to improve a given measurement procedure without increasing the test length and testing time, and deserves to be further explored.

TM033292

Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, Illinois, April, 2000

BEST COPY AVAILABLE

## 1. Purpose

The purpose of this study was to evaluate the expected precision increase in CAT scores by allowing items to contribute to multiple scores. Each item's contribution to a score was limited to the extent that the item discriminated among examinees with respect to that score. Two types of information that an item can provide are distinguished: "In-scale" and "out-of-scale". In-scale information is defined as information that an item provides for a composite trait to which it is specifically classified by a content review. Out-of-scale information is defined as information that an item provides for a composite trait other than the composite to which it is specifically classified.

It was hypothesized that out-of-scale information would increase the measurement precision of CAT scores. Two conditions were formed in computing information provided for simulees on measured traits. In the first condition, total information provided was computed by in-scale information alone. In the second condition, the total information provided was computed by in- and out-of-scale information together. Total information provided was compared for the two conditions with reference to the known true ability level of examinees. Associated bias and standard error measures were evaluated.

## 2. Background

In general, Multidimensional Item Response Theory (MIRT) models provide better model fit to test data than unidimensional IRT models do, and allow items to discriminate examinees along the dimensions involved in the measured trait. However, MIRT applications are not commonly used in modeling and scoring test data, primarily because MIRT applications require a great deal of item response data to adequately

estimate item and person parameters. MIRT model applications are recommended to test-developers to investigate the measurement process when unidimensional approaches are inappropriate (Ackerman, 1992, Miller & Hirsh, 1992, Wang, 1988).

Test developers continue to use unidimensional IRT models to analyze test data by relying on the robustness of IRT models and reporting as many scores as intended dimensions measured. There are two general approaches in fitting unidimensional IRT models to multidimensional response data. The first of the approaches assumes that response data of a test is "essentially unidimensional" if a major trait accounts for most of the examinee's performance to all items in a test. The residual performance is considered being influenced by minor traits that are nuisance dimensions on the measure obtained (Stout, 1990, Nandakumar, 1991).

The second approach assumes that response data of a test could be properly modeled fitting a unidimensional model if a single composite measure accounts for examinees' performance on a test (Wang, 1986, Reckase, Ackerman & Carlson, 1988, Ackerman, 1989). The "composite" usually refers to a linear combination of traits involved. The weights are related to dimensional strength of the underlying traits (e.g., the number of items measuring similar traits and the discrimination parameters of these items). However, they are not uniquely determined due to the "scale indeterminacy" property of IRT models, requiring reference composite scales to be determined (Wang, 1986). This approach aims to keep multidimensional structure in data, and, at the same time, allows unidimensional interpretations to be made with reference to a single trait composed of multiple traits.

In either case, however, to be justified in using a unidimensional model requires the principle of conditional independence to be approximately satisfied. That is to say

that one dominant dimension is present and that the dimensional structure of the response data is invariant across items and examinees.

### 3. Making Better Use of Information in Response Data

When there is evidence to believe that the test data at hand are not approximately unidimensional, a practical solution is to divide the scale into subscales, each calibrated separately and measuring a single trait composite, to avoid distortion in parameter estimates obtained through unidimensional models. Separate scores are reported for the subscales, forcing test developers increase the number of items administered to achieve acceptable reliabilities for these scores. Increasing testing time is a favorite of neither test-takers nor developers. Moreover, test developers are being asked to report an increasing number of scores for diagnostic purposes. And they are required to do so in an efficient way; that is, without increasing the number of items or testing time.

A few studies in the literature have proposed an approach for making better use of information provided by items in order to increase the reliabilities of scores reported without increasing total test length (Ackerman and Davey, 1991, Davey and Hirsh, 1991, Ackerman, 1994). The approach relies upon the idea that each item in a test measures multiple traits, and the traits are positively correlated. With this approach, it is feasible to think of unidimensional parameter estimates of a subtest of items as those items' unidimensional projections on the trait that the subtest is thought to measure. Therefore, unidimensional IRT models may be fitted to response data that are known to measure somewhat similar traits, thereby allowing as many unidimensional parameters to be estimated for each item as there are compound traits measured. Multiple sets of unidimensional item parameter estimates may be obtained through theoretical (Wang, 1986, Zhang and Wang, 1998) or empirical procedures (Ackerman and Davey, 1991,

Davey & Hirsh, 1991). Both procedures require pre-determination of a reference composite trait to fix the measurement scale. Then, items and persons represented by vectors are projected onto a reference composite scale to obtain unidimensional item and person parameters. Theoretical procedures can be used to derive unidimensional parameters, or unidimensional projections of multidimensional data structure, from a multidimensional model fitted to the response data. Empirical procedures, on the other hand, can be used to calibrate unidimensional parameters or unidimensional projections of multidimensional data structure from response data and test specifications.

### 4. Geometrical Representation of In- and Out-of-Scale Parameters

The solid arrows in Figure 1 represent two reading items as vectors in a two-dimensional space formed by the reading and math composites. The direction of the item vector indicates the direction in space the item best measures, while the length of the vector indicates how discriminating the item is in that direction of the space.

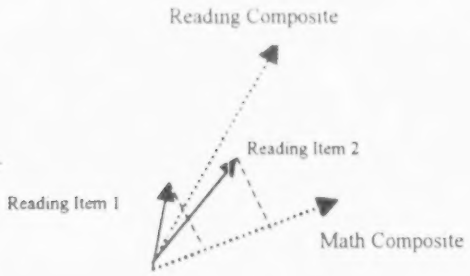


Figure 1. Geometrical representations of item and response vectors

Interpreted geometrically, an item discriminates with respect to a composite proportionally to the length of its projection on that composite. The dashed lines in the

figure show how the discriminations of the two reading items diminish when they are projected onto the math composite.

## 5. Method

Information provided for CAT simulees was compared under two conditions on two moderately correlated trait composites, math and reading comprehension. The Math and Reading Comprehension CATs were selected to illustrate a worst case scenario, as reading and math measures are expected to be as uncorrelated as any two cognitive measures that measurement specialist would want to use with out-of-scale information. The first condition used information provided by in-scale items alone, while the second condition used information provided by in- and out-of-scale items together in computing the total information provided for simulees at their true ability levels. True ability level of a simulee refers to a unidimensional approximation of its multidimensional true ability, where the unidimensional approximation was the 3PL ability with response probabilities of the MIRT Model (See Thompson, Davey, & Nering, 1998).

### Simulation

A realistic simulation procedure (Davey, Nering & Thompson, 1997) was used in generating the two-test battery of the CAT response data, math and reading comprehension. A high dimensional MIRT model was fit to real response data from a multiple-choice large-scale battery of tests used for college admission. The NOHARM computer program (Fraser & McDonald, 1988) was used to obtain multidimensional item parameters for the fitted model that were to serve as the true parameters in the simulation.

**Item pool simulation**

The CAT item pool consisted of 360 math and 360 reading comprehension items. Unidimensional item parameters were calibrated from 5000 examines simulated from the multidimensional model. In-scale and out-of-scale item parameters were estimated by BILOG (Mislevy & Bock, 1990) and the PIC computer program (Davey & Spray, 1999), fitting the 3PL IRT model to each test. The probability of a correct response in the 3PL model is given by

$$P_i(\theta) = c_i + (1 - c_i)\left[1 + e^{-D a_i(\theta - b_i)}\right]^{-1}, \qquad (1)$$

where *i* is the item administered, and $P_i(\bullet)$ is the 3PL model probability of a correct answer for the ith item for an examinee with ability θ (Hambleton & Swaminathan, 1985).

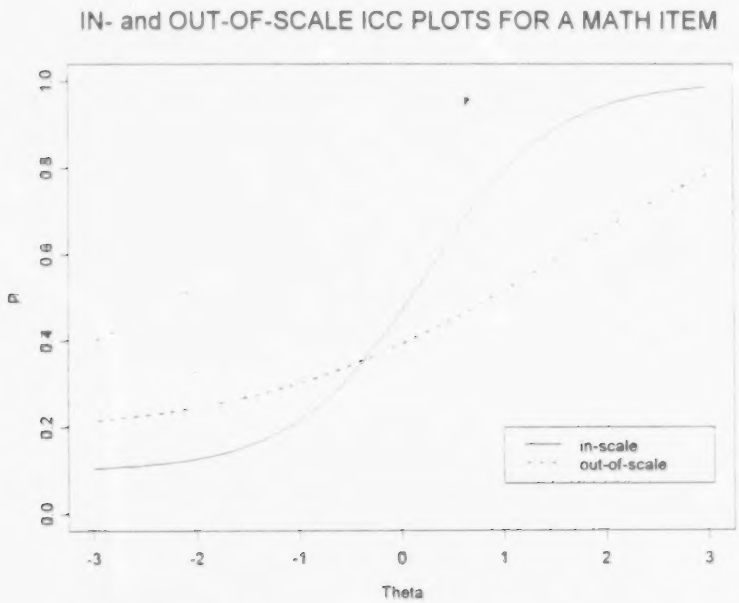


Figure 2. In and out-of-scale parameters for a math item

PIC uses the method of maximum likelihood to calibrate out-of-scale items one at a time for a composite trait. The in-scale item parameters are held constant in every PIC run to fix the scale, and out-of-scale items are calibrated individually, as the presence of other out-of-scale items would contaminate the scale.

For example, out-of-scale parameters of math items were calibrated one at a time with respect to the reading comprehension composite, which was defined by all the reading comprehension items in the CAT item pool. Each item had two sets of parameters, one in-scale and one out-of-scale. Figure 2 shows ICCs of a math item with respect to math (in-scale) and reading comprehension (out-of-scale) composites.

**CAT simulation**

A CAT data simulation program was developed to generate simulees from the fitted multidimensional model using the CAT item pool of 720 items. The simulated data included response data of the 20,000 simulees to both math and reading comprehension CATs with fixed lengths of 20 items. The estimated Pearson-Product moment correlation was 0.62 for true math ability and true reading comprehension ability of 20,000 simulees.

An item selection algorithm was used that maximized information for provisional ability estimates uniformly across the ability scale. Provisional ability estimates were obtained with the Expected A Priori (EAP) method. EAP is a Bayesian method in which information from the response pattern and information about the ability distribution (the mean of a prior distribution) are combined to obtain subsequent provisional ability estimates without restraining the prior distribution to be normal (Wang & Vispoel, 1998). The final ability estimates were obtained by the maximum likelihood method. The exposure rate was controlled with the Sympson-Hetter method (Davey & Parshall,

1995, Sympson & Hetter, 1985). With Sympson-Hetter exposure control, an item is not administered 100% of the time that it is selected as the optimal item, but only in 100$k_i$% Each item is assigned an exposure parameter, $k_i$, in the range of (0,1). Once the item that maximizes information given theta is selected, a random number from the uniform (0,1) distribution is generated. The random number is compared to the item exposure parameter of the optimal item, $k_i$, if $k_i$ is large the item is actually administered (Revuelta & Ponsoda, 1998, Meijer & Nering, 1999).

It should be noted here that the item selection algorithm of the CAT simulator only uses the in-scale item information functions in selecting the optimal item to be administered at each step.

**Dependent Variables**

Information provided, bias and standard error measures were calculated solely using the item parameters and evaluated only at true ability levels of simulees. True simulee ability was represented by unidimensional approximations of the multidimensional true ability.

**Information**

Two information functions were computed for each item, by

$$I(\theta) = \frac{(P_i')^2}{P_i Q_i} \qquad (2)$$

for the two conditions. For example each math item had two information functions, one with respect to the math composite, in-scale, and one with respect to the reading comprehension composite, out-of-scale.

In condition one, the total information was computed for each simulee's true math and reading comprehension ability level by summing the information provided by 20 in-scale items that each examinee was administered. In condition two, the information provided for the true math and reading ability levels of each examinee were computed by summing the information provided by 40 items, 20 of which were out-of-scale items and represented by a unidimensional approximation. Equation 3 shows the test information equation

$$I(\theta_j) = \sum_{i=1}^{n} \left( \frac{(P_i)^2}{P_i Q_i} \right). \qquad (3)$$

where total information provided for jth simulee with ability $\theta_j$ was computed by summing the information provided by i=(1,...,n) items (Hambleton & Swaminathan, 1985). It must be noted here that the n is 20 and 40 for the first condition and the second condition, respectively.

**Standard Error**

The standard error of the ability estimate was computed by

$$SE(\theta_j) = \frac{1}{\sqrt{I(\theta_j)}} \qquad (4)$$

for each condition on the reading comprehension and math composites.

**Bias**

Equation 5 shows the theoretical bias function derived by Lord (1983) and generalized by Samejima (1993a, 1993b).

$$Bias(\theta) \cong \frac{D}{I^2} \sum_{i=1}^{n} a_i I_i (P^* - 0.5) \quad \text{where} \quad P^* = \frac{P_i - c_i}{1 - c_i} \qquad (5)$$

According to Equation 5, bias will be close to zero when all items are targeted at the examine's true ability level, $\theta$. Bias will be negative when the true ability level is higher than the average item difficulty level, and will be positive when true ability level is lower than the average item difficulty level.

## 6. Results

Information provided for the true ability of simulees was compared for the two conditions with the associated observed standard error and bias. The results are summarized for math and reading comprehension reference composites for simulees grouped according to their true abilities. 11 groups were formed along the ability continuum in the range of (-1.8,1.8) with increments of 0.4. Table 1 and Table 2 show the information provided by the Math and Reading composites for simulees at their true abilities (unidimensional approximations of their multidimensional abilities).

Information provided for simulees on the reading composite increased approximately 17% when math items were allowed to contribute to the reading comprehension composite, represented with their reading comprehension scale item parameters (out-of-scale parameters). Information provided for the math composite increased approximately 16% when reading comprehension items were allowed to contribute to the math composite, represented with their math-scale item parameters (out-of-scale parameters).

A 17% gain in information corresponds to a 3.4 increase in the number of items administered (17% of 20). Even though this does not seem to be a big gain, this information is "free" in the sense that response data at hand already included both tests. Thus, if both tests could be shortened by three items, the total administered would be 34 instead of 40.

## Table 1. Simulee Groups Summaries for Math Composite

| | Simulee Groups* | Math In-Scale | | | Math In & Out-of-Scale Total | | |
|---|---|---|---|---|---|---|---|
| Statistic | True Math Ability | Info. | SD | Bias | Info. | SD | Bias |
| Mean | -1.4<theta<=-1.8 | 2.645425 | 0.397788 | -0.120948 | 3.319941 | 0.312274 | -0.084929 |
| *Variance* | 0.012427 | 0.311235 | 0.011058 | 0.004612 | 0.367899 | 0.004152 | 0.001143 |
| Mean | -1.0<theta<=-1.4 | 3.775837 | 0.276218 | -0.071113 | 4.637967 | 0.222435 | -0.052320 |
| *Variance* | 0.013161 | 0.544807 | 0.003850 | 0.001601 | 0.624843 | 0.001759 | 0.000564 |
| Mean | -1.0<theta<=-0.6 | 5.200953 | 0.198298 | -0.040661 | 6.289044 | 0.162695 | -0.031207 |
| *Variance* | 0.013293 | 0.784116 | 0.001408 | 0.000713 | 0.872959 | 0.000652 | 0.000272 |
| Mean | -0.6<theta<=-0.2 | 6.827584 | 0.150673 | -0.023760 | 8.185468 | 0.124753 | -0.018534 |
| *Variance* | 0.013116 | 1.216837 | 0.000798 | 0.000481 | 1.343938 | 0.000359 | 0.000181 |
| Mean | -0.2<theta<=-0.2 | 8.684694 | 0.118244 | -0.012065 | 10.308054 | 0.098974 | -0.009651 |
| *Variance* | 0.01359 | 1.86751 | 0.000422 | 0.000319 | 2.023302 | 0.000213 | 0.000140 |
| Mean | 0.2<theta<=0.6 | 10.428054 | 0.098197 | -0.001575 | 12.275323 | 0.082900 | -0.001526 |
| *Variance* | 0.013134 | 2.342721 | 0.000265 | 0.000250 | 2.458913 | 0.000133 | 0.000115 |
| Mean | 0.6<theta<=1.0 | 11.778176 | 0.086658 | 0.007080 | 13.784278 | 0.073643 | 0.005099 |
| *Variance* | 0.013356 | 2.503644 | 0.000185 | 0.000261 | 2.582239 | 0.000009 | 0.000123 |
| Mean | 1.0<theta<=1.4 | 12.882337 | 0.079506 | 0.013044 | 14.986672 | 0.067880 | 0.009644 |
| *Variance* | 0.013228 | 3.11854 | 0.000214 | 0.000321 | 3.204647 | 0.000101 | 0.000140 |
| Mean | 1.0<theta<=1.4 | 13.563158 | 0.075745 | 0.021545 | 15.705057 | 0.064840 | 0.016069 |
| *Variance* | 0.012829 | 3.061097 | 0.000328 | 0.000383 | 3.110929 | 0.000130 | 0.000130 |

*The range includes 18,597 simulees out of 20,000 simulated. Number of simulees included in each group are 826, 1442, 2128, 2781, 3402, 3228, 2473, 1552, 765

## Table 2. Simulee Groups Summaries for Reading Composite Composite

| | Simulee Groups | Reading C. In-Scale | | | Read.C. In & Out-of-Scale Total | | |
|---|---|---|---|---|---|---|---|
| Statistic | True Math Ability | Info. | SD | Bias | Info. | SD | Bias |
| Mean | -1.4<theta<=-1.8 | 2.147006 | 0.703801 | -0.128059 | 2.782265 | 0.608652 | -0.071811 |
| *Variance* | 0.01361 | 0.217838 | 0.023766 | 0.100223 | 0.239098 | 0.005098 | 0.001551 |
| Mean | -1.0<theta<=-1.4 | 2.74045 | 0.621573 | -0.091652 | 3.514062 | 0.540933 | -0.047526 |
| *Variance* | 0.013099 | 0.32517 | 0.021726 | 0.213937 | 0.359279 | 0.003634 | 0.001073 |
| Mean | -1.0<theta<=-0.6 | 3.38633 | 0.55404 | -0.054693 | 4.298542 | 0.487619 | -0.032995 |
| *Variance* | 0.013188 | 0.402223 | 0.007229 | 0.017891 | 0.440335 | 0.002192 | 0.000802 |
| Mean | -0.6<theta<=-0.2 | 4.324165 | 0.48646 | -0.040445 | 5.394166 | 0.433905 | -0.027098 |
| *Variance* | 0.013143 | 0.501007 | 0.002169 | 0.002283 | 0.551535 | 0.001071 | 0.000581 |
| Mean | -0.2<theta<=-0.2 | 6.11853 | 0.41001 | -0.027938 | 7.342935 | 0.372877 | -0.019947 |
| *Variance* | 0.013385 | 1.42535 | 0.001598 | 0.000786 | 1.520314 | 0.00096 | 0.000323 |
| Mean | 0.2<theta<=0.6 | 8.678873 | 0.344978 | -0.006477 | 10.028314 | 0.319663 | -0.004945 |
| *Variance* | 0.01317 | 3.144029 | 0.001365 | 0.000431 | 3.224592 | 0.000864 | 0.000216 |
| Mean | 0.6<theta<=1.0 | 10.594378 | 0.311295 | 0.014906 | 12.036946 | 0.291163 | 0.011678 |
| *Variance* | 0.01282 | 3.591243 | 0.000964 | 0.000247 | 3.632109 | 0.00063 | 0.000129 |
| Mean | 1.0<theta<=1.4 | 10.331641 | 0.314059 | 0.034227 | 11.824009 | 0.292921 | 0.026647 |
| *Variance* | 0.012933 | 2.570564 | 0.000659 | 0.000264 | 2.606721 | 0.000432 | 0.000133 |
| Mean | 1.0<theta<=1.4 | 8.552314 | 0.344652 | 0.053062 | 10.042797 | 0.317423 | 0.039763 |
| *Variance* | 0.013277 | 1.443017 | 0.000676 | 0.000421 | 1.522382 | 0.000419 | 0.000173 |

*The range includes 18,541 simulees out of 20,000 simulated. Number of simulees included in each group are 838, 1449, 2223, 2917, 3188, 3077, 2418, 1563, 868



BEST COPY AVAILABLE

Figure 3 and Figure 4 show the plotted average information provided for the 11 simulee groups for two conditions; in- and total (in- and out-of-scale together). In-scale information provided takes a curvilinear form with increasing true ability for the reading comprehension composite, while it preserves its linear form for the math composite. The math CAT provided more information in general for simulees than the reading comprehension CAT did.

Mutual information contributed by out of scale items alone to the scales seem to increase slightly with increasing ability level of simulees, more so for the math composite (out-of-scale information increases 42% and 31% from the lowest to the highest true ability group for math and reading comprehension composites, respectively). However, the increase does not seem to be of a practical importance for the range considered in this study (e.i. the increase in magnitude is approximately 2.0 for the math and 0.9 for the reading composites).

Figure 5 and Figure 6 respectively present plotted average standard error and bias associated with the information provided in two conditions across the ability composites. The standard error observed decreased approximately 17% and 10% percent for math and reading comprehension composites on average. The decrease observed in the computed average bias for theta groups was approximately 23% for the math composite and 31% for the reading comprehension composite. The decrease in the standard error was higher in magnitude for the reading comprehension, while the decrease in bias was higher in magnitude for the math composite. The decrease pattern was similar for the two CATs across the ability scales, taking into account that observed standard error values computed for the reading comprehension were higher due to relatively less information provided by the reading comprehension CAT.

Overall, information provided for simulees was higher for condition 2 with decreased bias and standard error terms. The contribution of the reading comprehension and the math composite items to the other composite was approximately the same in magnitude. The math composite had larger standard errors and smaller bias values when compared to the reading comprehension composite. However, the standard error decrease was more drastic for the math composite, while bias decrease was more drastic for the reading comprehension composite.

Further study is needed to compare the two conditions on a bias measure that is concerned with how close an examinee's final estimated ability is to his/her approximated true ability. With out-of-scale information this type of bias is expected to increase, not to decrease. However, because ML theta estimates are biased outward in general, the additional bias due to out-of-scale information may be expected to show a counter-balancing effect.

## 7. Discussion

An increasing number of test users are demanding more and better diagnostic information from the tests their students take. It would seem that only very long tests can be expected to provide the number of highly reliable subscores some test users desire. The challenge to test developers then is to increase the number of subscores reported while simultaneously keeping test length to an acceptable level.

The performance of out-of-scale information procedure seems promising in CAT settings. 17% information gain found would be a minimum that one would expect by using out-of-scale information. Measures that correlated higher than the math and reading measures would yield more substantial information gains.

The results of this study show that out-of-scale information can be used to improve a given measurement procedure without increasing the test length and testing

time, and deserves to be further explored. The performance of the proposed application must be studied for a wide variety of conditions (i.e., dimensional strength and inter-correlations of composites). For example, examinees's final ability estimates could be obtained with and without incorporating out-of-scale information. With out-of-scale information, examinee's final estimated ability is expected to be biased toward the mean ability estimate. The ultimate aim would be to delineate the conditions under which out-of-scale information would insure better measurement procedures with out-of-scale information than without it.

**References**


Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. Applied Psychological Measurement, v. 13, n. 2, pp. 113-127.

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. Journal of Educational Measurement, v. 29, n. 1, pp. 67-01.

Ackerman, T. A. (1994). Creating a test information profile for a two-dimensional latent space. Applied Psychological Measurement, v. 18, pp. 257- 275.

Ackerman, T. A. & Davey, T. C. (1991, April). Concurrent adaptive measurement of multiple abilities. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Davey, T. C. & Hirsh, T. M. (1991, April). Examinee discrimination and the measurement properties of multidimensional tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Davey, T. C., Nering, M. L. & Thompson, T. (1997). Realistic simulation of item response data. ACT Research Report, series 97-4.

Davey, T. C., Parshall, C. G. (1995, April). New algorithms for item selection and exposure control with Computerized Adaptive Testing. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Davey, T. C., & Spray, J. (1999). PIC: Pretest item calibration computer program. {Computer Program}. Iowa City IA: ACT, Inc.

Fraser, C. & McDonald, R. P. (1988). NOHARM: An IBM PC computer program for Fitting both unidimensional and multidimensional normal ogive models of latent trait theory {Computer Program}. Center for Behavioral Studies, The University of New England, Armidale, New South Wales, Australia.

Hambleton, R. K., & Swaminathan, H. (1985). Item Response Theory: Principles and Application. Boston: Kluwer-Nijhoff Pub.

Lord, E. M. (1983). Unbiased estimators of ability estimators, their variance, and of their parrale-forms reliability. Psychometrika, v.48, pp. 233-245.

Meijer, R. R. & Nering, M. L. (1999). Computerized Adaptive Testing: overview and Introduction. Applied Psychological Measurement, v. 23, n. 3, pp.187-194.

Miller, T. R. & Hirsh, T. M. (1992). Cluster analysis of angular data in applications of multidimensional item response theory. Applied Psychological Measurement, v. 5, n. 3, pp.193-211.

Misleyv, R. J. & Bock, R. D. (1990). BILOG 3: Item analysis and test scoring with binary logistic models. Moorisville IN: Scientific Software.

Nandakumar, R. (1991). Traditional versus essential dimensionality. Journal of Educational Measurement, v. 28, n. 2, pp. 361-373.

Reckase, M. D., Ackerman, T. A. & Carlson, J. E. (1988). Building a unidimensional Test using multidimensional items. Journal of Educational Measurement, v. 25 n. 3, pp. 193-203.

Revuelta, J., Ponsoda, V. (1998). A comparison of item exposure control methods in Computer Adaptive

Testing. Journal of Educational Measurement, v. 35, n. 4, pp. 311-327.

Samejima, F. (1993a). An approximation for the bias function of the maximum likelihood estimate of the latent variable for the general case where the item responses are discrete. Psychometrika, v. 58, pp. 119-138.

Samejima, F. (1993b). The bias function of the maximum likelihood estimate of ability for dichotomous response level. Psychometrika, v. 58, pp. 195-209.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimates. Psychometrika, v. 55, pp. 293-325.

Sympson, J. B., Hetter, R. D. (1985). Controlling item exposure rates in Computerized Adaptive Testing. Proceedings of the 27th Annual Meeting of the Military Testing Association (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

Thompson, T. D., Davey, T. & Nering, M. (1998, April). Constructing adaptive tests to parallel conventional programs. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Wang, M. (1986, April). Fitting a unidimensional model to multidimensional response data. Paper presented at the ONR Contractors Conference, Gatlinburg, TN.

Wang, M. (1988, April). Measurement bias in the application of a unidimensional model to multidimensional item response data. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Wang, T. Vispoel, W. P. (1998). Properties of ability estimation methods in Computerized Adaptive Testing. Journal of Educational Measurement, v. 35, n. 2, pp. 109,135.

Zhang, J., Wang, M. (1998). Relating reported scores latent traits in a multidimensional test. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

FIGURE 3: Information provided for the Math composite

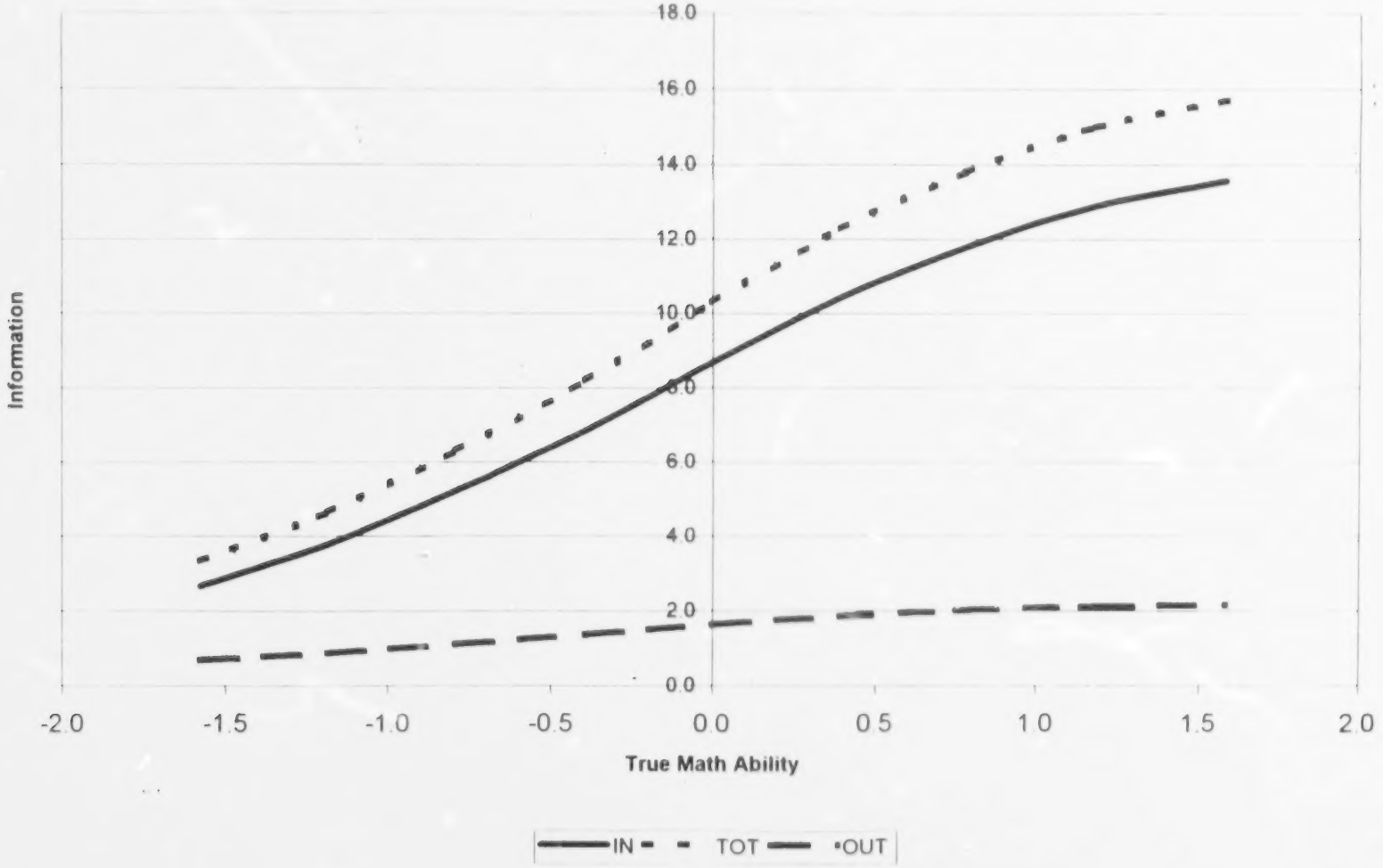

FIGURE 4: Information Provided for reading Comprehension Composite

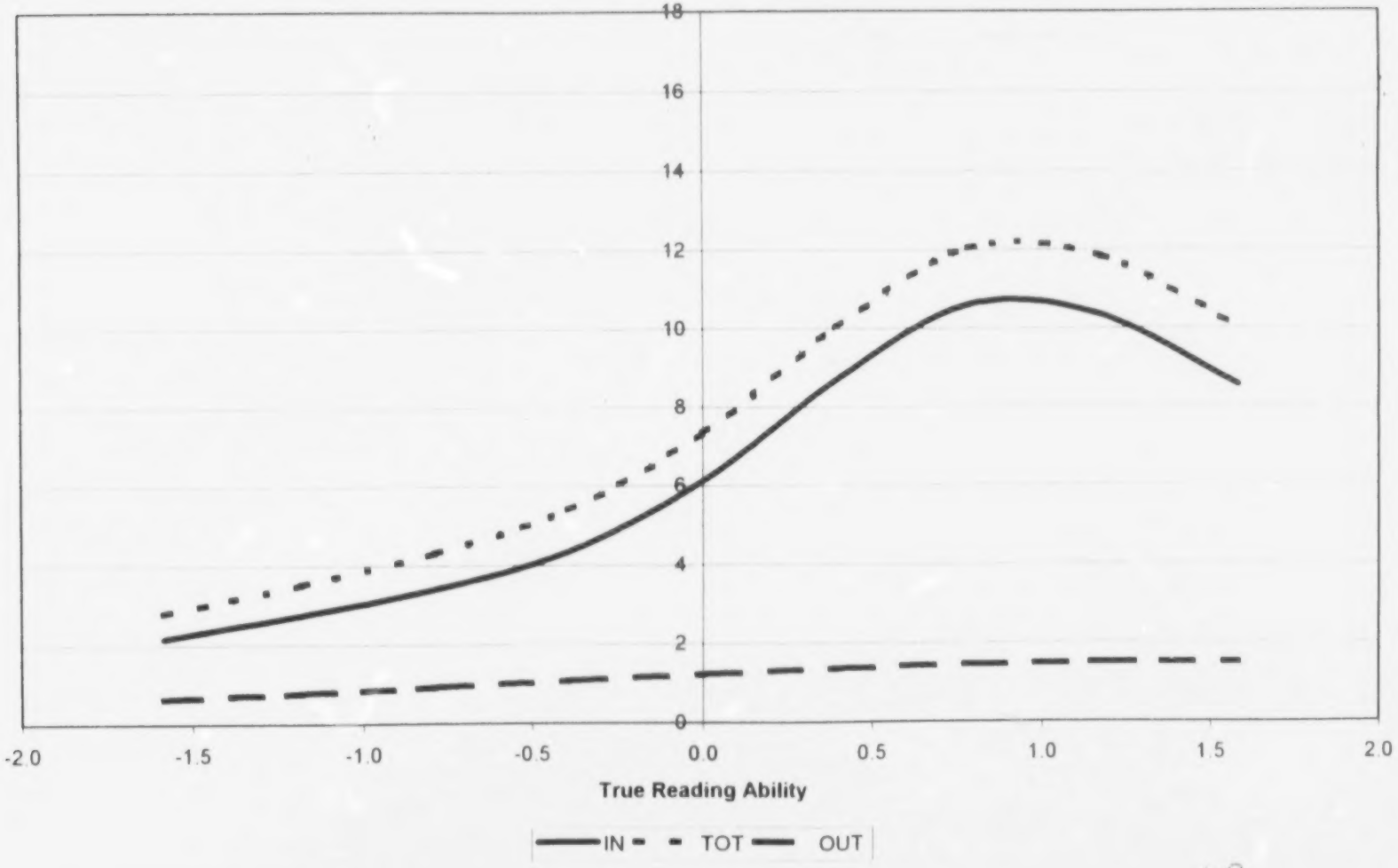

FIGURE 5: Bias versus Standard Error
Math Composite

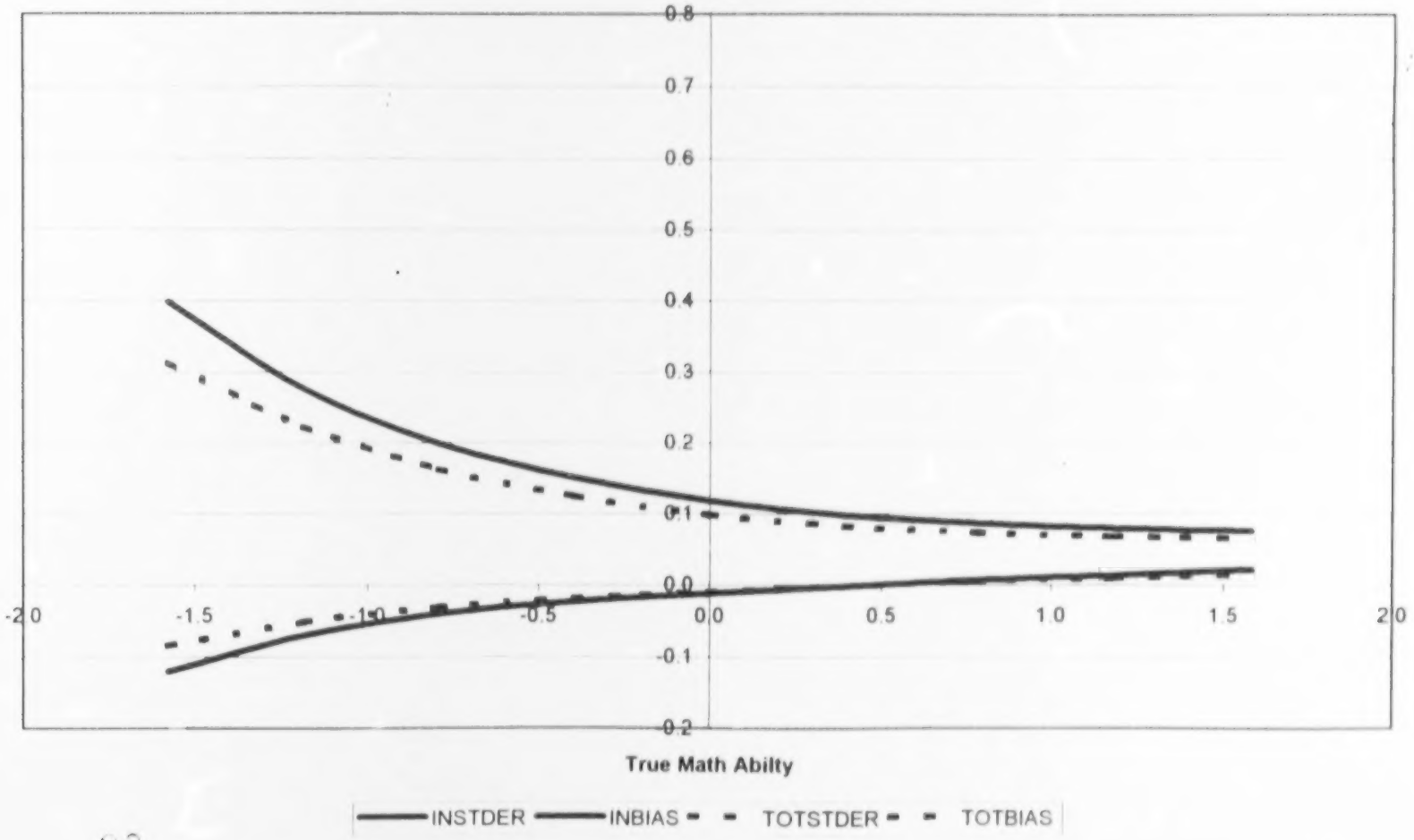

**FIGURE 6: Standard Error versus Bias**
**Reading Comprehension Composite**